Forecasting Candy Production Index in the US

1. About the Data

About

Sweets, chocolates, and candy are universally enjoyed. In the US, there are holidays themed around giving candy! All this consumption first needs production. The dataset below shows the monthly production of candy in the US. The industrial production index measures the actual output of all relevant establishments in the United States, regardless of ownership, but not those in U.S. territories.

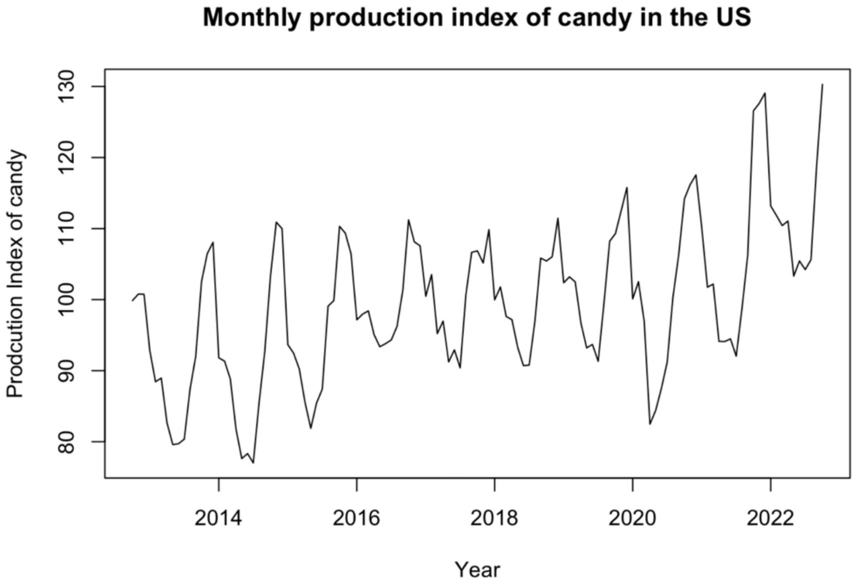
Data Source

Link: https://fred.stlouisfed.org/series/IPG3113N

Data Dictionary

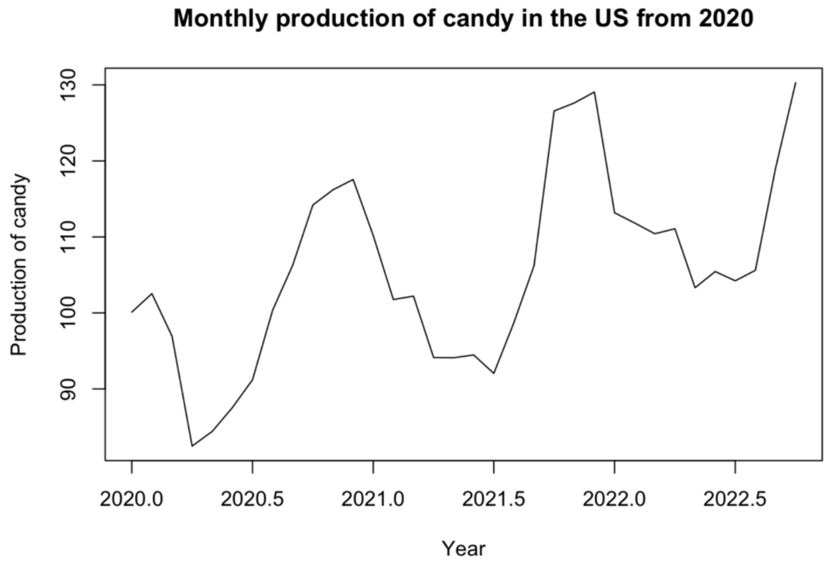
Date: Year, Month, and Date during with the data was recorded
IPG3113N: Production Index for Candy in the US

Plot & Inferences

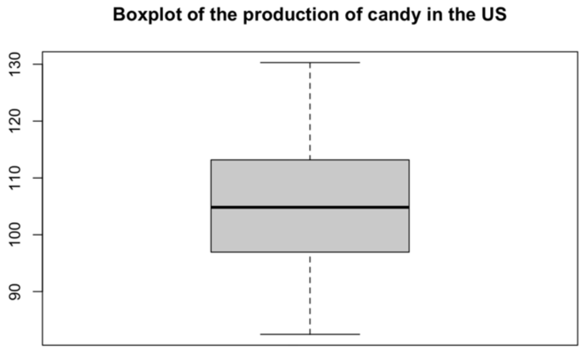


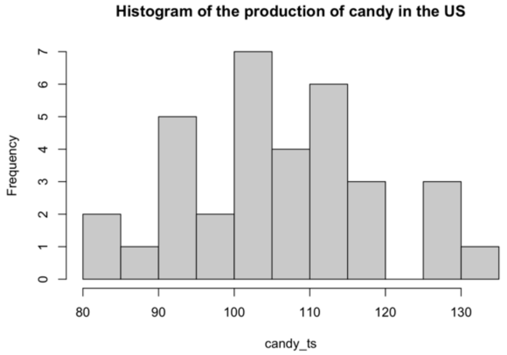
- The data from 2012 has seasonal variation and is peaking every November and December every year.
- This is because of the holiday season every year that has Thanksgiving and Christmas.
- However, from 2020, the data has an increasing trend and seasonal component.

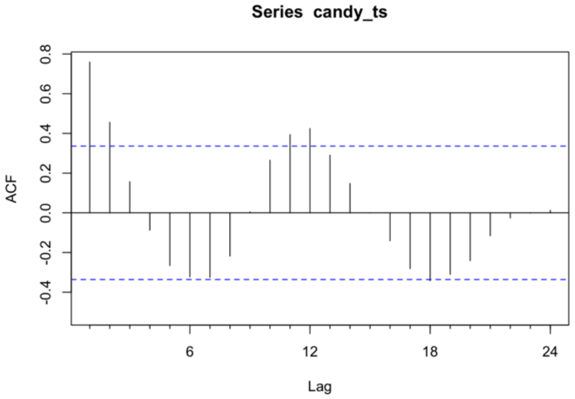
- To explore this idea more, we consider a window starting from 2020 and considering two years of data will be good enough for a proper forecast.



- Considering the window function, the plot has both trend and seasonality.
- Forecasting this data will be more accurate as it is the recent data, and there is a high chance that the future data will have the same trend and seasonality.
- Further analysis of the data will be done considering this data set.

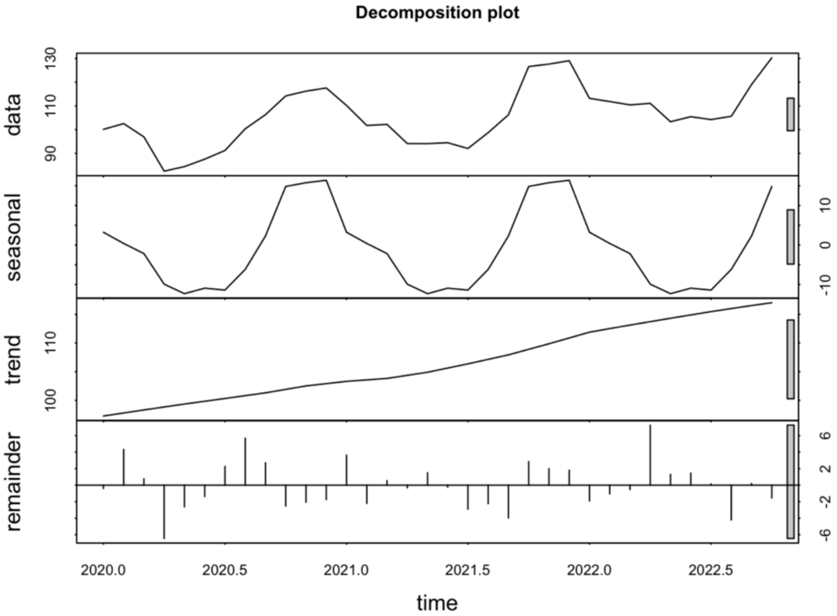






- The boxplot shows that there are no outliers in the data.
- The data has a mean of 105.63 and doesn't look to have a proper normal distribution.
- The median is in between the 1st and 3rd quartile and is not specifically towards one of them.
- From the summary, we can also see that the median value is less than the mean for the time series.
- This means that the data is right-skewed. This can be justified by seeing the histogram above as well.
- The ACF plot shows a strong trend and seasonality in the data. The trend can be inferred based on the number of lines crossing the confidence interval.
- The strong seasonality can be inferred based on the wavy nature of the Acf plot, and the seasonality period is 12 months. We can see a peak and dip every six months simultaneously.
- We can observe the same thing in the plot as well.

2. Decomposition



- The decomposition plot shows a trend and seasonality in the time series data.
- Based on this analysis, we can develop a question and hypothesis to start our forecast.
- The decomposition is additive because, as trend increases, we do not see any increase in the seasonality. The seasonality appears to be the same throughout.

3. Question and Hypothesis

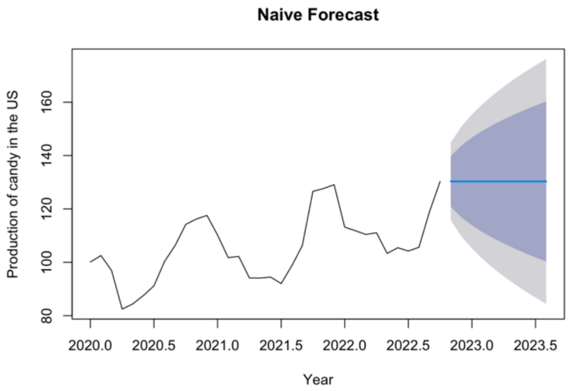
Question

What will be the best method to forecast the given time series data?

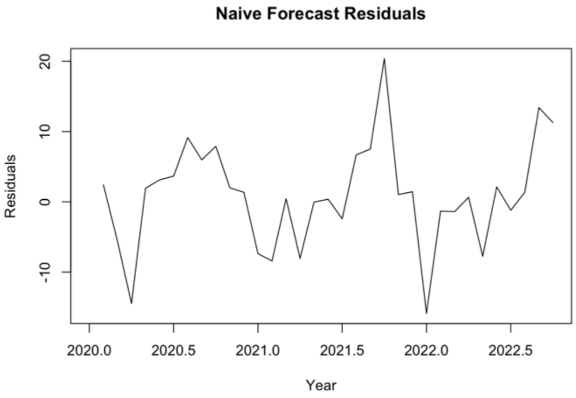
Hypothesis

- The modern ANOVA method might give us the best forecast for time series by expanding our knowledge from previous forecasting techniques.
- We can check this hypothesis based on the accuracy of each model that we can check below.

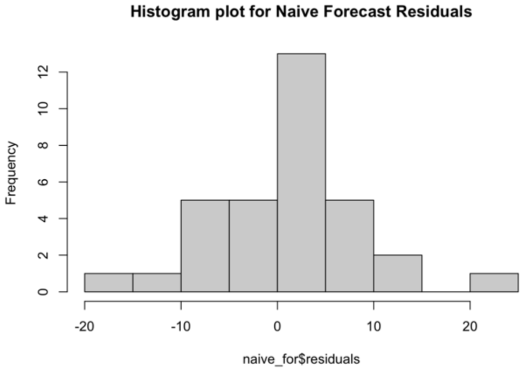
4. Forecast - Naïve Method



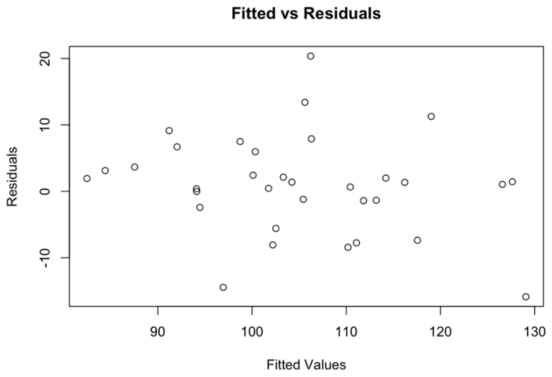
Naive Forecast



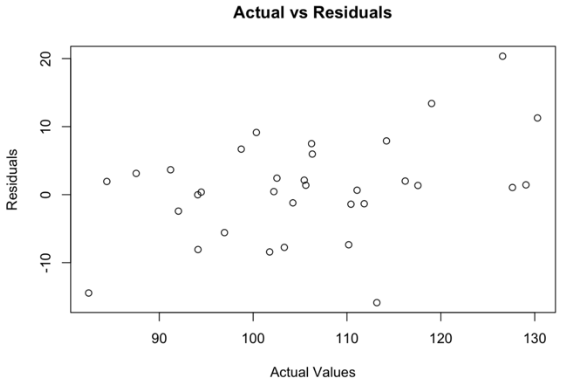
Naive Forecast Residuals



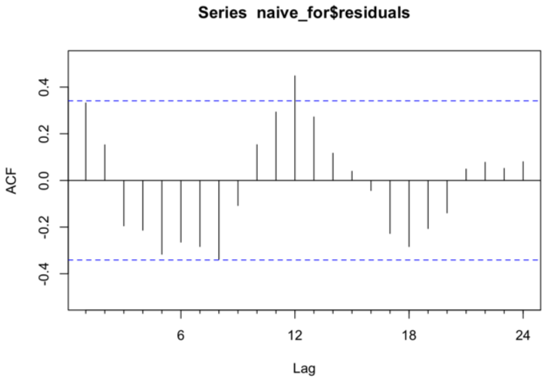
Histogram plot for Naive Forecast Residuals



Fitted vs Residuals



Actual vs Residuals



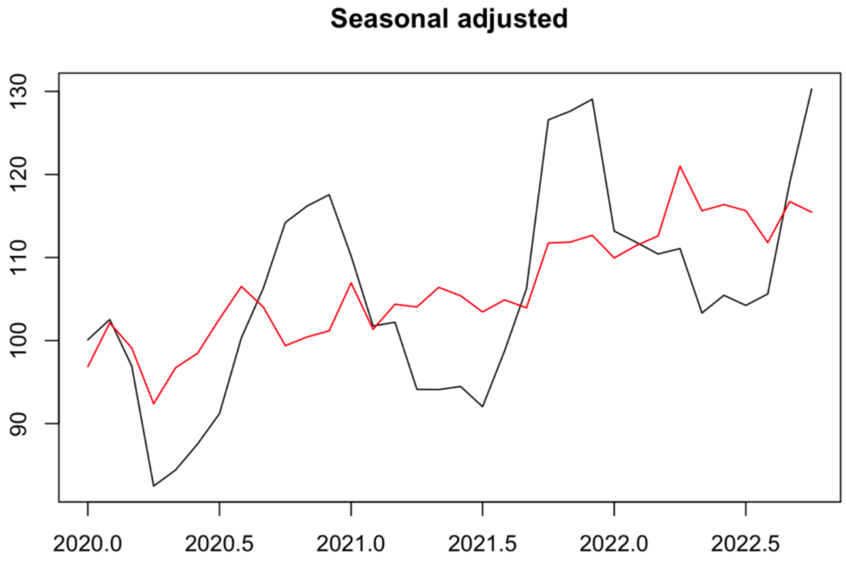
Accuracy

```
accuracy(naive_for)
```

```
##                     ME     RMSE      MAE       MPE     MAPE      MASE      ACF1
## Training set 0.9147333 7.399605 5.399739 0.5619459 5.090241 0.6740498 0.3322229
```

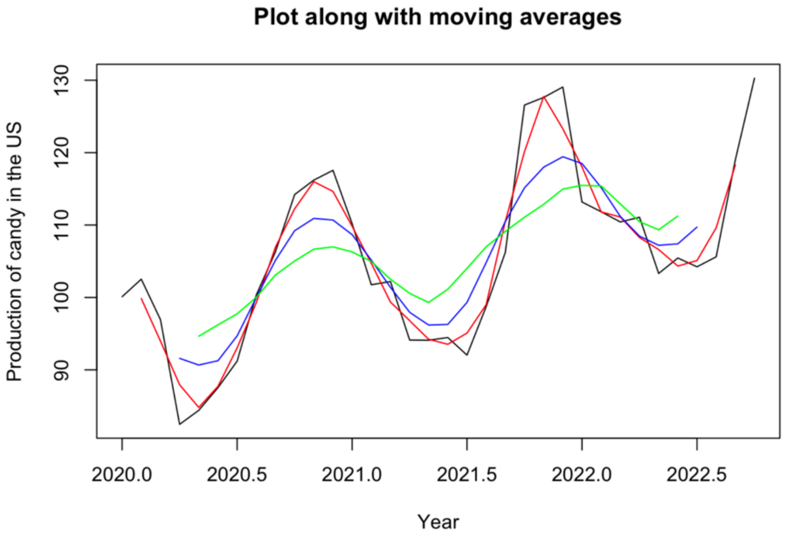
- The histogram suggests a somewhat normal distribution appearance.
- Data lacks a mean of zero and exhibits skewness, indicating bias.
- Residuals show randomness until 2022, then an increasing trend emerges, suggesting unaccounted-for factors.
- The fitted vs. Residuals plot is relatively random but contains three outliers.
- The actual vs. residual plot displays a cone shape, indicating increasing residuals over time.
- The Autocorrelation Function (Acf) of residuals shows both trend and seasonality.
- The model assessment suggests the need to consider additional variables or alternative forecasting techniques to address issues with residuals.

5. Seasonally adjusted plots



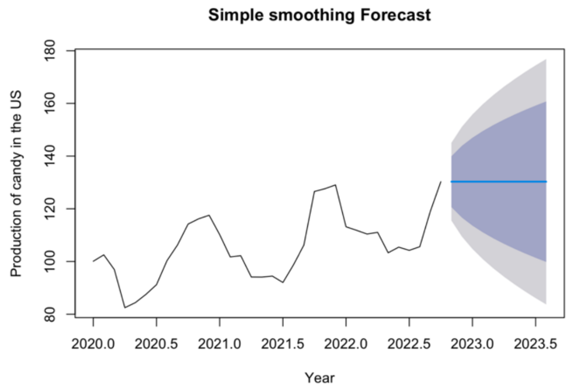
- The seasonality has significant fluctuations in the value of the time series.
- This is expected, as the data showed strong seasonality in the ACF plot.

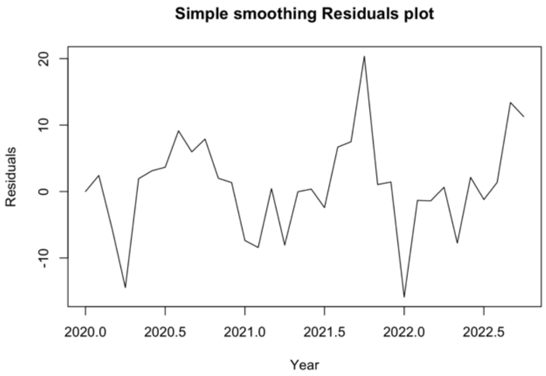
6. Simple Moving Averages

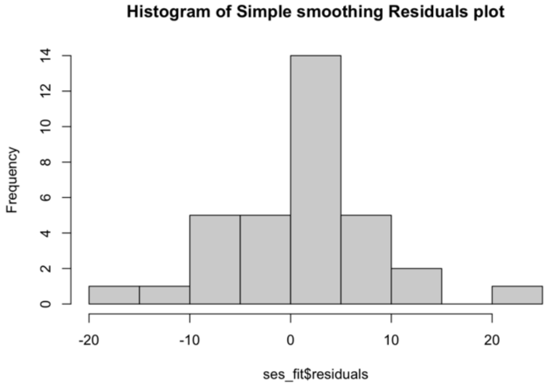


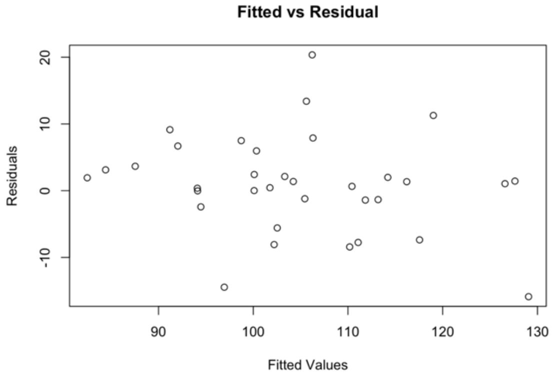
- The plots show that the higher the order we consider, the smoother the moving average curve in the plot.
- It can be seen that the Green line above is the smoothest compared to the Blue or Red lines.
- The Red line (order 3) gives the most actual data compared to the other two. The higher order averages smoother the plot and does not give the actual values.

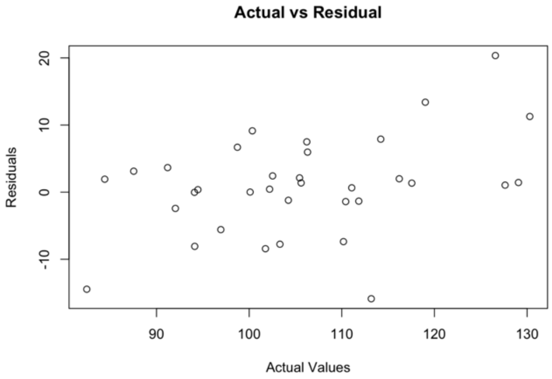
7. Forecast - Simple Smoothing

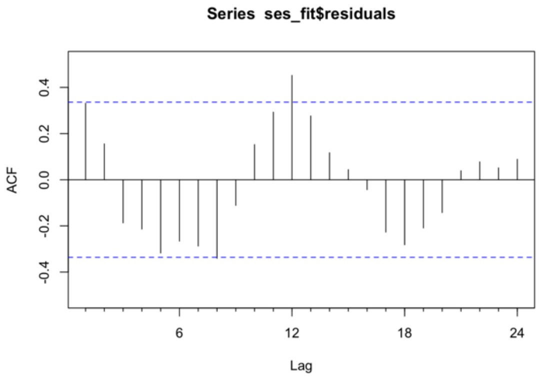












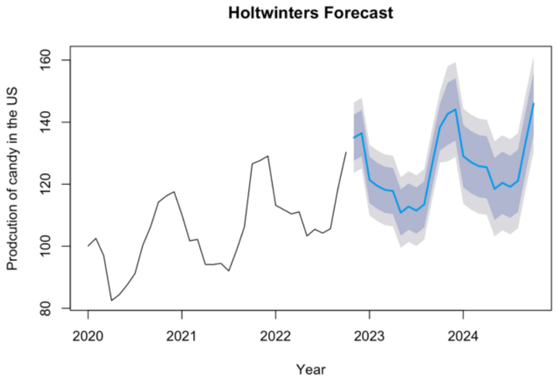
Accuracy

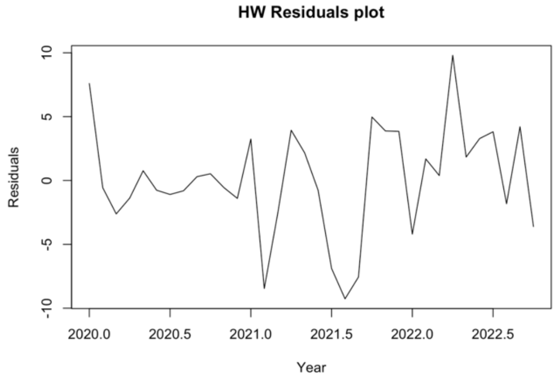
```
accuracy(ses_fit)
```

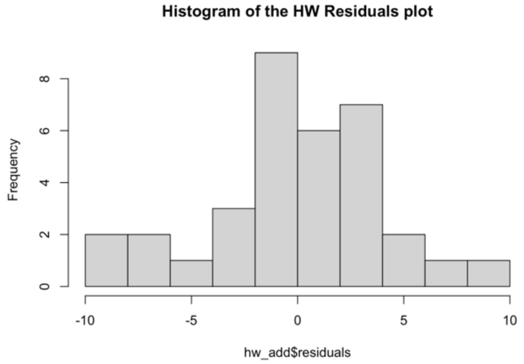
```
##                      ME    RMSE      MAE       MPE     MAPE      MASE      ACF1
## Training set 0.8882407 7.29022 5.241508 0.5457879 4.941071 0.6542978 0.3312411
```

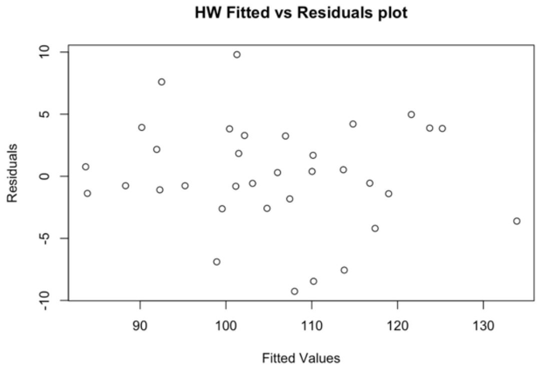
- Residuals show randomness until 2022, after which they exhibit an increasing trend, indicating the omission of crucial factors.
- The histogram suggests a non-zero mean and skewness, signifying bias in the data despite its normal distribution appearance.
- While the Fitted vs. Residuals plot appears acceptable with a mean around zero, it contains three outliers.
- The Actual vs. Residuals plot displays a concerning cone shape, indicating increasing residuals over time and the omission of influential variables.
- In conclusion, the current method has limitations in capturing certain factors influencing the data. Exploring other forecasting techniques, like Holt-Winters, is advisable to improve the model's accuracy and address the observed issues with residuals.

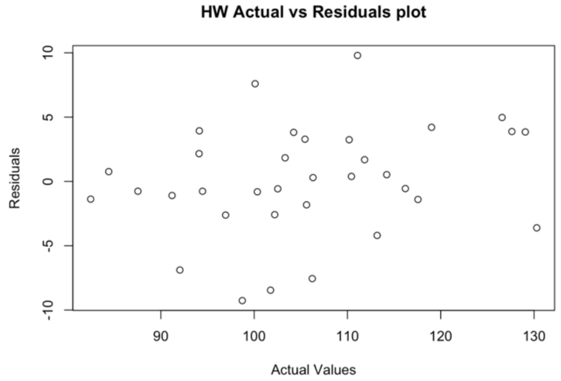
8. Forecast – Holt-Winters

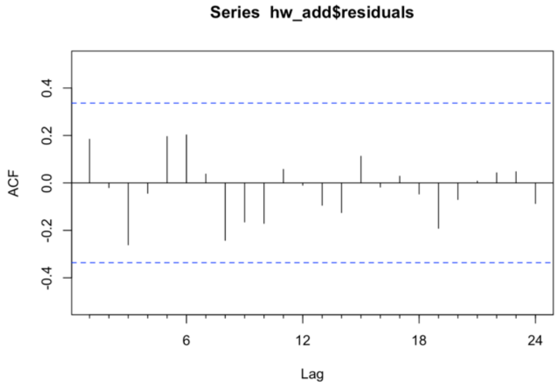


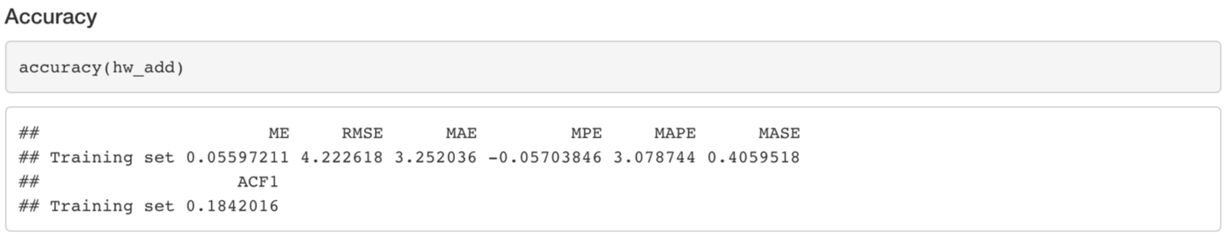












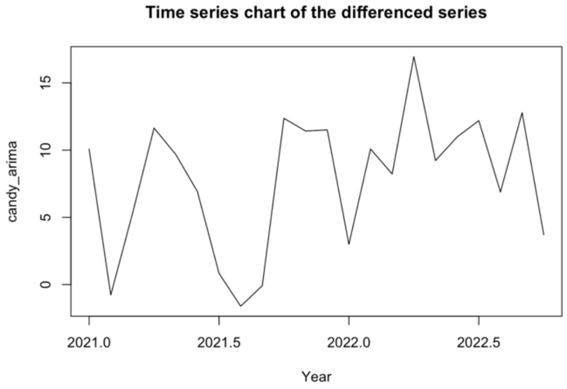
Accuracy

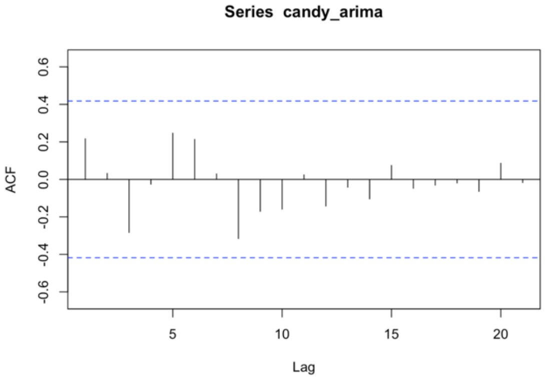
```
accuracy(hw_add)
```

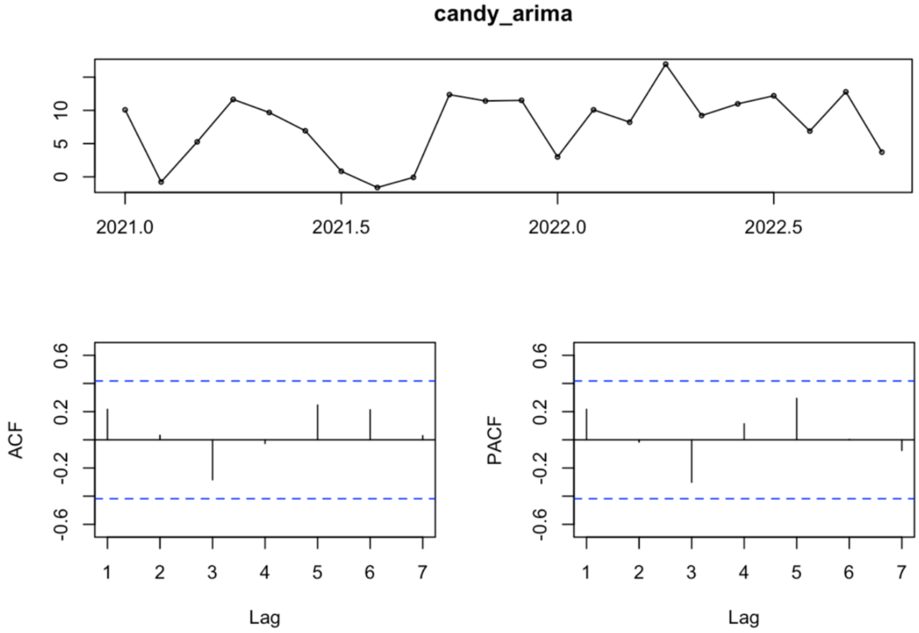
```
##                        ME     RMSE      MAE         MPE     MAPE      MASE
## Training set 0.05597211 4.222618 3.252036 -0.05703846 3.078744 0.4059518
##                   ACF1
## Training set 0.1842016
```

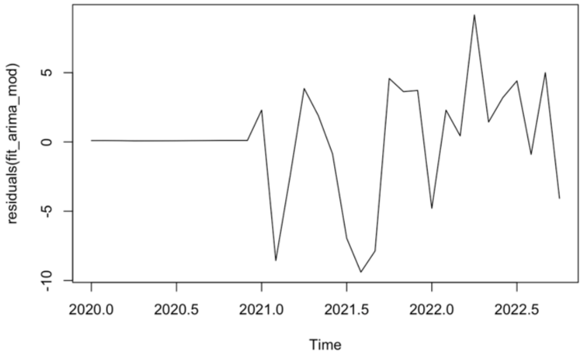
- Residuals are random with a near-zero mean, indicating a well-fitted model.
- The histogram is approximately normally distributed, though the data has a bias.
- Fitted vs. Residuals and Actual vs. Residuals plots show random patterns but also reveal outliers.
- ACF plot suggests white noise behaviour in the residuals, a positive sign for forecasting.
- Overall, the Holt-Winters model performs well, but there is room for improvement, possibly through ARIMA modelling, given the identified outliers and data bias.

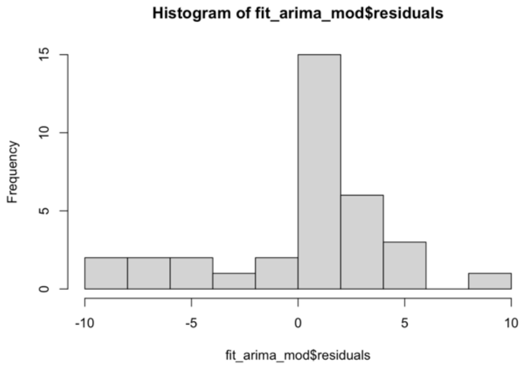
9. ARIMA

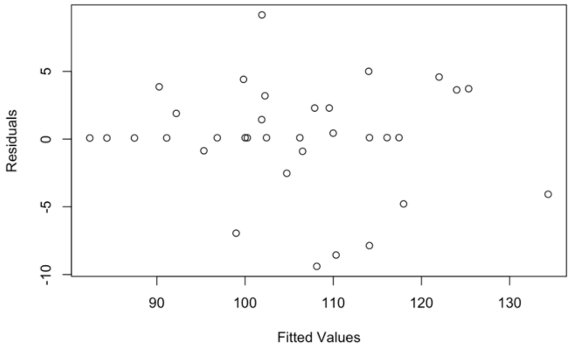

Time series chart of the differenced series
candy_arima
0
5
10
15
2021.0
2021.5
2022.0
2022.5
Year

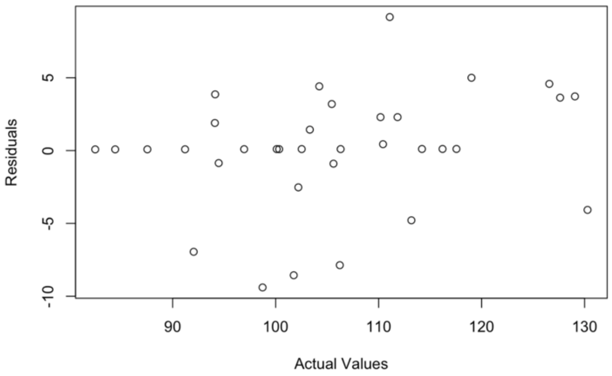

Series candy_arima
ACF
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
5
10
15
20
Lag

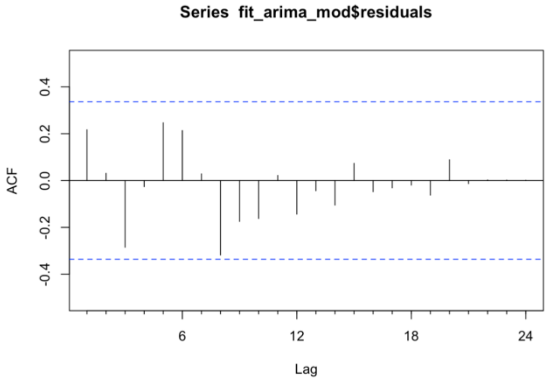

candy_arima
0
5
10
2021.0
2021.5
2022.0
2022.5
ACF
-0.6
-0.2
0.2
0.6
1
2
3
4
5
6
7
Lag
PACF
-0.6
-0.2
0.2
0.6
1
2
3
4
5
6
7
Lag

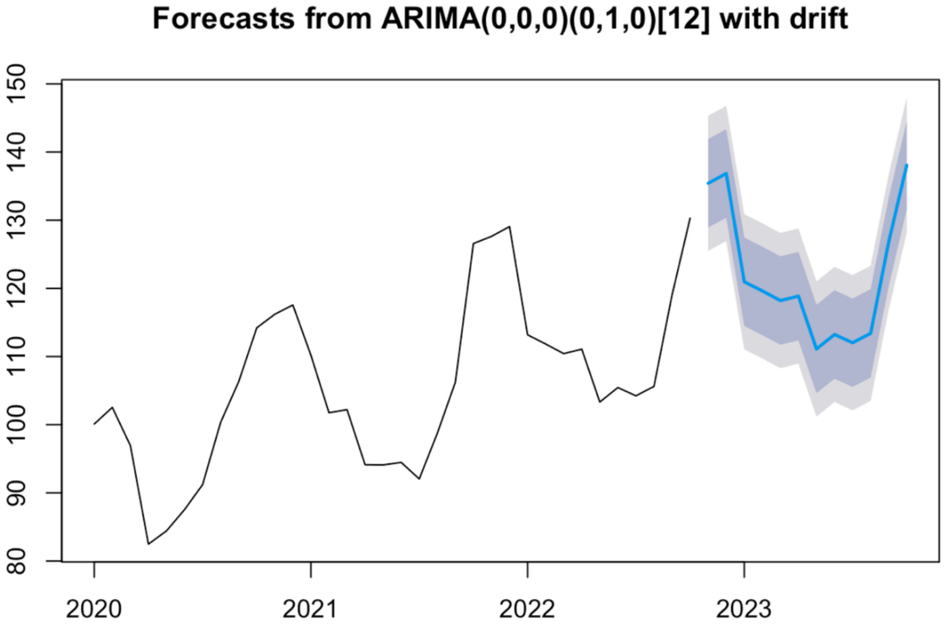

residuals(fit_arima_mod)
-10
-5
0
5
2020.0
2020.5
2021.0
2021.5
2022.0
2022.5
Time


Histogram of fit_arima_mod$residuals
Frequency
0
5
10
15
-10
-5
0
5
10
fit_arima_mod$residuals









- The time series data is non-stationary, exhibiting both trend and seasonality.
- Seasonal differencing is performed to address seasonality, potentially taking care of trend differencing.
- ACF and PACF plots indicate that the potential ARIMA models could include terms like ARIMA(0,1,0)(0,1,0) or variations thereof.

- Model selection explores values other than 0 for p, q, P, and Q to consider alternative models based on AIC and BIC criteria.
- Residual analysis reveals randomness in residuals and a mean near zero, indicating a well-fitted model.
- While some fluctuations are observed in residual plots, they do not suggest a growing residual pattern.
- The histogram of residuals appears to be normally distributed but lacks a mean of zero, implying data bias.
- Fitted vs. Residuals and Actual vs. Residuals plots display randomness with a mean around zero, yet a few outliers are noted.
- ACF plot indicates residuals behaving like white noise, a positive sign for forecasting accuracy.
- ME and RMSE values are low, and ARIMA modelling emerges as the best forecasting method, supported by both accuracy measures and residual analysis.

10. Conclusion

Based on the analysis of the Candy Production Index data in the US, here is the conclusion:

The Candy Production Index data in the US exhibits clear seasonality and an increasing trend, with peaks occurring every November and December due to the holiday season. Starting from 2020, there has been a noticeable upward trend in production. The data is right-skewed and does not follow a normal distribution.

Several forecasting methods were explored, including the Naïve Method, Simple Moving Averages, Simple Smoothing, Holt-Winters, and ARIMA models. Each technique had its strengths and limitations.

- The Naïve Method showed increasing residuals after 2022, indicating unaccounted-for factors, suggesting the need for more sophisticated models.
- Simple Moving Averages demonstrated that higher-order averages smoothed the data but did not provide actual values.
- Simple Smoothing showed increasing residuals over time and omitted influential variables, indicating the need for alternative forecasting techniques.
- Holt-Winters performed well, with random residuals and a near-zero mean, but it still had outliers and data bias.
- ARIMA modelling was found to be the most suitable forecasting method. It addressed seasonality and trend through differencing and provided well-fitted residuals with low ME and RMSE values. The ACF plot indicated that residuals behaved like white noise, a positive sign for forecasting accuracy.

In conclusion, ARIMA modelling is recommended as the best method for forecasting the Candy Production Index data in the US. However, it's important to continually monitor and adjust the model as new data becomes available, as external factors can impact candy production in unforeseen ways. Additionally, exploring further model improvements and considering additional variables may enhance forecasting accuracy.